Diffusion d'information dans les réseaux sociaux

Mots-clés: Réseaux sociaux (on line), diffusion, viralité, communautés, topologie, polarisation, fake-news, motifs, Twitter.

Contexte:

La compréhension des mécanismes de circulation/diffusion/propagation des discours, des opinions, des fake-news, des rumeurs dans les réseaux sociaux numériques devient un enjeu de société. La viralité et la détection des robots ont été étudiées lors d'événements majeurs tels que les élections présidentielles américaines en 2016 [Kollanyi 2016], le brexit [Howard 2016] ou encore pendant l'élection présidentielle française de 2017 [Ferrara 2017]. Les algorithmes développés utilisent des techniques de *machine learning* en agrégeant plusieurs centaines de critères, ils permettent une détection assez fiable sur des événements précis mais il ne sont pas généralisables, et nécessitent une phase d'apprentissage coûteuse et n'apportent un éclairage que sur certains aspects de l'étude.

La compréhension de ces mécanismes soulève plusieurs questions :

- Quel est l'impact de la topologie du réseau dans les phénomènes de diffusion d'un message viral (et son importance par rapport au contenu des messages) ?
- Quels types de relations dans un réseau sont les plus à même d'amplifier la diffusion ?
- Comment les communautés et surtout les communautés polarisées participent à la diffusion ?
- Quel est le rôle des robots, leur comportement ?
- Comment les personnes influentes ou les *leaders* d'opinion agissent dans cette diffusion ?
- Est-il possible de généraliser les mécanismes observés c'est-à-dire quels patterns, structures ou processus types participent ou conditionnent la diffusion à grande échelle ?

Sujet:

Pour répondre à ces questions plusieurs algorithmes ont été développés et validés sur des jeux de données spécifiques pour détecter des communautés [Drif 2014, Orman 2012], des utilisateurs influents [Riquelme 2016, Ibnoulouafi 2018], des robots [Ferrara 2016], des événements [Atefeh 2015], des messages viraux, rumeurs [Sela 2017, Zubiaga 2018, Hoang 2011], etc.

Les problématiques abordées dans cette thèse sont les suivantes :

- la première est relative à la modélisation des données et a pour but de développer un modèle de données [Leclercq 2018] à partir de la notion de réseau complexe multi-couches [DeDomenico 2013, Kivela 2014] qui permettra de représenter et d'exploiter les différentes relations en leur donnant une sémantique appropriée. Par exemple Twitter, par la richesse des relations issues des opérateurs (follow, retweet, mention,

etc.) génère des réseaux complexes dont la sémantique est cachée. Il constitue donc un bon terrain d'expérimentation.

- la seconde concerne la combinaison d'algorithmes afin de répondre aux interrogations des chercheurs en sciences sociales. Cet enjeu est majeur car il permet de comprendre des processus complexes, d'élaborer des modèles et de les tester sur les données. A titre d'exemple l'étude de l'influence ne peut pas être dissociée de la notion de communauté [Weng 2013, Kumar 2018, Gupta 2016, Gupta 2015] et des frontières de communautés, il est est de même pour la propagation des messages viraux qui peuvent se diffuser à l'intérieur d'une communauté (déjà acquise), ou se propager à l'extérieur, voire mẽme à modifier la structure communautaire. Une des pistes prometteuses est d'utiliser les techniques de *graph embeding* [Hongyun 2017, Goyal 2017]

D'un point de vue expérimental, le sujet abordera dans un premier temps la mise en place de différents algorithmes d'analyse, en s'appuyant sur des jeux de données déjà collectés par l'équipe1 pour :

1°) mesurer l'audience/l'impact d'une thématique ou d'un événement [Atefeh 2015]. Un point à aborder est la viralité du discours et l'impact des robots dans la propagation du discours [Ferrara 2016] ;

2°) montrer l'existence de communautés dans lesquels le discours circule. Ces communautés doivent être caractérisées [Basaille 2018, Jebabli 2014, Jebabli 2015] pour mettre en avant leur particularité/singularité. Mais les communautés s'imbriquent, s'intersectent et il faut aussi étudier leur influence réciproque sur la circulation du discours ;

3°) étudier l'influence d'une communauté vers une autre et les personnes qui font les liens (élasticité des frontières) et par conséquent revoir la notion d'influenceurs [Azaza 2016, Jebabli 2015a], de leader d'opinion en fonction de la circulation des informations.

Dans un second temps, à partir des résultats expérimentaux, il conviendra de développer un ou plusieurs modèles de diffusion et de les tester sur de nouveaux jeux de données en exploitant leur aspect prédictif pour valider leur aspect explicatif [Jebabli 2018].

D'un point de vue plus théorique, une des problématiques abordées traite de l'adaptation et de la combinaison d'algorithmes traditionnels aux réseaux complexes multi-couches et l'extension des techniques de *graph embeding*, les preuves des propriétés des algorithmes proposés devront être abordées comme par exemple la convergence, les mesures qualités par rapport à une vérité de terrain, etc.

Cette thèse est centrée sur les aspects fondamentaux de modèle de données et d'outils d'analyse pour réseaux complexes mais s'appuie sur des collaborations institutionnelles établies depuis 2013 avec les laboratoires de l'Université de Bourgogne TIL et CIMEOS en Sciences Humaines et Sociales au travers de projets interdisciplinaires TEE 2014, TEP 2017, PEPS CNRS MOMIS.

Profil du candidat:

Le/la candidat(e) doit donc pouvoir mener à bien une recherche innovante et de grande

qualité. Il/elle devra développer des recherches visant à: (i) traiter les problèmes théoriques fondamentaux et (ii) concevoir des modèles et des algorithmes pouvant être utilisés pour comprendre les processus de diffusion dans les réseaux multicouches.

Il/elle doit avoir de bonnes connaissances en mathématiques appliquées, algèbre linéaire, statistiques et informatique (algorithmique).

Les autres exigences sont:

Bonne maîtrise des langages de programmation tels que R, Python pour l'analyse de données et C ++ ou équivalent pour les simulations informatiques, l'accès à la base de données et le stockage.
Expérience - ou intérêt à développer - des techniques efficaces d'analyse de données à grande échelle.
Très bonne maîtrise de l'anglais (oral et écrit) et excellentes compétences en communication.
La curiosité, l'autonomie, l'intégrité et la créativité sont des qualités souhaitées

Formation et compétences requises:

Le/la candidat(e) doit être titulaire d'une maîtrise (ou équivalent) dans le domaine de l'informatique, des mathématiques appliquées ou d'une discipline connexe, obtenue avec une très bonne note finale (avec une note moyenne de B ou supérieure).

Equipe de Recherche:

L'équipe Science des Données du Laboratoire d'Informatique de Bourgogne est fortement axée sur la recherche quantitative, mais appliquée avec des compétences reconnues en Système d'information, IA, Big Data et réseaux Complexes.
Elle offre :
- La possibilité d'achever un doctorat dans le domaine de la science des réseaux et du Big Data en faisant appel à des outils de systèmes complexes et à la modélisation des interactions socio-économiques.
- Un travail large et indépendant au sein d'une équipe dynamique dans une atmosphère de travail positive.
- Un programme de développement de carrière complet (participation à des écoles d'été, conférences, etc.).

Directeur de thèse : Hocine Cherifi
Co-encadrants : Eric Leclercq et Marinette Savonnet

Comment postuler

Les candidatures doivent être envoyées par courrier électronique, avec les éléments suivants:

- un CV ,
- les relevés des notes universitaires
- une déclaration d'intérêts (une page, maximum)

- les noms et coordonnées de deux référents.

Les demandes de renseignements supplémentaires peuvent être envoyées aux contacts (adresse ci-dessous).
Adressez votre correspondance avec comme sujet : **Thèse diffusion d'information dans les réseaux sociaux**

Contacts :
Hocine Cherifi - Hocine.Cherifi@u-bourgogne.fr
Marinette Savonnet – Marinette.Savonnet@u-bourgogne.fr
Eric Leclercq – Eric.Leclercq@u-bourgogne.fr

Laboratoire d'Informatique de Bourgogne (LIB) – EA 7534
Équipe Science de Données
Université de Bourgogne
9, Avenue Alain Savary
21078 Dijon

Bibliographie (en gras publications des membres de l'équipe):

[Atefeh 2015] Atefeh, Farzindar, and Wael Khreich. "A survey of techniques for event detection in twitter." Computational Intelligence, 31.1 (2015): 132-164.
[Azaza 2016] Azaza, Lobna, Kirgizov, Sergey, Savonnet, Marinette, *et al.* Information fusion-based approach for studying influence on twitter using belief theory. *Computational Social Networks*, 2016, vol. 3, no 1, p. 5-25.
[Basaille 2018] Basaille Ian, Plateforme pour la gestion des données issues des réseaux sociaux dans le cadre de la gestion de la relation client, Thèse de doctorat, Université de Bourgogne, 2018.
[Jebabli 2018] Jebabli M., Cherifi H., Cherifi C., Hammouda A., "Community detection algorithm evaluation with ground-truth data, Physica A: Statistical Mechanics and its Applications 492, 651-706 Elsevier 2018
[Jebabli 2015] Jebabli M., Cherifi H., Cherifi C., Hammouda A., "Overlapping community detection versus ground-truth in AMAZON co-purchasing network" in 11th International Conference on Signal Image Technology & Internet-Based Systems, Proceedings of IEEE 2015
[Jebabli 2015a] Jebabli M., Cherifi H., Cherifi C., Hammouda A., "User and group networks on YouTube: A comparative analysis, in 12th International Conference on Computer Systems and Applications (AICCSA), Proceedings of IEEE 2015
[Jebabli 2014] Jebabli M., Cherifi H., Hammouda A.,"Overlapping Community Structure in Co-authorship Networks: A Case Study," in 7th International Conference on u- and e-Service, Science and Technology (UNESST), Proceedings of IEEE pp.26,29, 2014
[Leclercq 2018] Leclercq, Eric, Savonnet, Marinette, "Modèle tensoriel pour l'entreposage et l'analyse des réseaux sociaux – Application à l'étude de la viralité sur Twitter", INFORSID 2018, à paraître.
[DeDomenico 2013] De Domenico, Manlio, et al. "Mathematical formulation of multilayer networks." *Physical Review X* 3.4 (2013): 041022.
[Drif 2014] Drif, Ahlem, and Abdallah Boukerram. "Taxonomy and survey of community

discovery methods in complex networks." *International Journal of Computer Science and Engineering Survey* 5.4 (2014): 1.
[Ferrara 2016] Ferrara, Emilio, et al. "The rise of social bots." *Communications of the ACM* 59.7 (2016): 96-104.
[Ferrara 2017] Ferrara, Emilio. "Disinformation and social bot operations in the run up to the 2017 French presidential election." (2017).
[Goyal 2017] P Goyal, E Ferrara, "Graph embedding techniques, applications, and performance: A survey", arXiv preprint arXiv:1705.02801, (2017).
[Gupta 2016] Gupta N., Singh A., Cherifi H. , " Centrality measures for networks with community structure", Physica A: Statistical Mechanics and its Applications 452, 46-59, Elsevier 2016
[Gupta 2015] Gupta N., Singh A., Cherifi H., "Community-based Immunization Strategies for Epidemic Control " in 7th International Conference on Communication Systems and Networks, Proceedings of IEEE , 2015
[Hoang 2011] Hoang, Tuan-Anh, et al. "On modeling virality of twitter content." *International Conference on Asian Digital Libraries*. Springer, Berlin, Heidelberg, 2011.
[Hongyun 2017] Cai, Hongyun, Vincent W. Zheng, and Kevin Chen-Chuan Chang. "A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications." *arXiv preprint arXiv:1709.07604* (2017).
[Howard 2016] Howard, Philip N., and Bence Kollanyi. "Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum." *Browser Download This Paper* (2016).
[Ibnoulouafi 2018] Ibnoulouafi A. El Hassouni M., Cherifi, "M-Centrality: Identifying key nodes based on global position and local degree variation" In revision for Journal of Statistical Mechanics: Theory and Experiment
[Kivela 2014] Kivelä, Mikko, et al. "Multilayer networks." Journal of complex networks 2.3 (2014): 203-271.
[Kollanyi 2016] Kollanyi, Bence, Philip N. Howard, and Samuel C. Woolley. "Bots and automation over Twitter during the first US Presidential debate." *Comprop data memo* 1 (2016): 1-4.
[Kumar 2018] Kumar M. Singh A., Cherifi H., " An efficient Immunization Strategy using Overlapping Nodes and its neighborhoods" to appear in the proceedings of the Web Conference 2018 (WWW18), Lyon, France
[Orman 2012] Orman G.K, Labatut V., and Cherifi H., "Comparative evaluation of community detection algorithms: a topological approach", Journal of Statistical Mechanics: Theory and Experiment, P08001, august 2012.
[Riquelme 2016] Riquelme, Fabián, and Pablo González-Cantergiani. "Measuring user influence on Twitter: A survey." Information Processing & Management 52.5 (2016): 949-975.
[Sela 2017] Sela, Alon, et al. "Increasing the Flow of Rumors in Social Networks by Spreading Groups." arXiv preprint arXiv:1704.02095 (2017).
[Varol 2018] Varol, Onur. Analyzing Social Big Data to Study Online Discourse and Its Manipulation. Diss. Indiana University, 2017.
[Weng 2013] Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. "Virality prediction and community structure in social networks." Scientific reports 3 (2013): 2522.
[Zubiaga 2018] Zubiaga, Arkaitz, et al. "Detection and Resolution of Rumours in Social Media: A Survey." ACM Computing Surveys (CSUR) 51.2 (2018): 32.